



May 22, 2012

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-10

MEMORANDUM FOR David C. Whitford
Chief, Decennial Statistical Studies Division

From: Patrick J. Cantwell *(Signed)*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Douglas Olson
Richard Griffin
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Report: Aspects
of Modeling

This report is one of twelve documents providing estimation results from the 2010 Census Coverage Measurement program. This report describes how modeling was used to create the estimates used for census coverage evaluation.

For more information, contact Doug Olson on (301) 763-9290.

Attachments

cc:
DSSD CCM Contacts List

Census Coverage Measurement Estimation Report

Aspects of Modeling

Prepared by
Douglas Olson
Richard Griffin

Decennial Statistical Studies Division

Table of Contents

Executive Summary	1
1. Introduction.....	2
2. Methods.....	2
2.1 Dual System Estimation.....	2
2.2 Correlation Bias Adjustment.....	3
2.3 Statistical Testing.....	3
2.4 Software	3
2.5 Model Fit (risk function) by Log-Likelihood	4
2.6 Parameter Choice by Wald Likelihood tests	4
2.7 Testing for Over-Parameterization by Cross-validation	4
2.8 Continuous Variable Transformations by Residual Examination and Plots.....	4
3. Limitations	4
3.1 Sampling Error.....	4
3.2 Nonsampling Error.....	5
3.3 Prefer to Use Same Covariates in Each Rate Model.....	5
3.4 Standard Techniques and Software.....	5
3.5 Modeling Assumption about Additivity	5
3.6 Characteristics.....	5
4. Discussion of Results for Person Estimation	6
4.1 Are All These Characteristics of Value?	6
4.2 Changes to the Existing Variables	8
4.3 Transformation of the Participation Rate.....	9
4.4 Choice of Interactions	11
4.5 Choice of Model for the Data-Defined rate	12
4.6 Testing for Model Fit	12
4.7 Final Model Check	13
5. Discussion of Results for Housing Unit Estimation	14
5.1 Are All These Characteristics of Value?	15
5.2 Transformation of Continuous Variables.....	15
5.3 Choice of Interactions	16

5.4 Testing for Model Fit	17
5.5 Final Model Check.....	17
6. Conclusions	18
References	19
Attachment 1: Variable Definitions	21
Attachment 2: Housing Unit Rate Residual Plots.....	25

Executive Summary

This report summarizes the statistical modeling diagnostics used to develop the logistic regression models selected to produce the estimates of net coverage error for persons and housing units for the 2010 Census Coverage Measurement.

For person estimation three logistic regression models were developed:

- a model to predict the probability of having Data-defined status in the census
- a model to predict the probability of a census enumeration being a Correct Enumeration
- a model to predict Match rate, the probability of a person being captured in the census

Model diagnostics used to define and select appropriate predictor variables and which interactions of these to put in the models were

- Wald Chi-Square tests
- model fit assessment by log-likelihood
- checking for over-parameterization by cross-validation
- determining if transformations (for example taking the square root of a rate) were needed for continuous predictor variables by examinations of residual plots

The analysis resulted in selecting the same set of main effects and interactions for all three logistic regression models. The predictor variables selected were

- Race/Hispanic Origin domains
- Tenure
- Age/Sex groups
- Region of the country
- Metropolitan Statistical Area Size by Type of Enumeration Area
- Presence of Spouse in Household
- Relationship to Householder
- Tract-level Census Participation Rates
- Bilingual and Replacement Questionnaire Mailing Areas

Housing unit modeling requires the estimation of rates for Correct Enumeration and Match to the census. The same model-building procedures used in person estimation resulted in the selection of a slightly different choice of interactions for the two models. The characteristics used in housing unit modeling are

- Structure type and size of the dwelling
- Occupancy and tenure
- Region of the country
- Metropolitan Statistical Area size by Type of Enumeration Area (TEA)
- Measures of the number of address list changes in the neighborhood near to Census Day
- Bilingual and Replacement Questionnaire Mailing Areas

1. Introduction

The purpose of the 2010 Census Coverage Measurement (CCM) program is to evaluate coverage of the 2010 Census to aid in improving future censuses. The CCM is designed to measure the census coverage of housing units and persons, excluding group quarters and persons residing in group quarters. The CCM sample design is a probability sample of 170,000 housing units. Remote areas of Alaska are out of scope for the CCM. The CCM provides estimates of net coverage estimates and components of census coverage by using a post-enumeration survey.

Prior to the 2010 CCM, the Census Bureau used a post-stratification approach to dual system estimation (DSE) to evaluate net coverage. This approach limits the number of independent variables that can be used because each variable added can crudely be thought of as cutting the post-stratum sample size in half. This is due to the implicit estimation of the many high-order interactions across variables used in the post-stratification. Logistic regression modeling methods were employed to estimate net coverage for the first time in the 2010 CCM. Logistic regression allows the addition of variables as main effects without the need to estimate parameters associated with higher-order interactions. As with post-stratification, the use of logistic regression allows many choices for the final model to be used for estimation.

This report summarizes the statistical modeling diagnostics used to develop the logistic regression models selected to produce the estimates of net coverage for persons and housing units for the 2010 CCM.

2. Methods

In this section, we first discuss briefly the estimation method used in generating the net coverage estimates for persons and housing units. Details of the logistic regression modeling diagnostics used to select the models are then provided. For more details on the CCM estimation methodology, see Mule (2008).

2.1 *Dual System Estimation*

Since the 1950 census, the Census Bureau has conducted post-enumeration evaluations to estimate the size of error in census counts for areas and demographic groups and to use the information to improve future census processes. The post-enumeration survey for 2010, called the 2010 CCM survey, relies on DSE that requires two independent systems of measurement. The Population Sample, P sample, and the Enumeration Sample, E sample, have traditionally defined the samples for DSE. The P sample and the E sample are intended to measure the same housing unit and household population. However, the P-sample operations are conducted independent of the census. The E sample consists of census housing units and person enumerations in housing units in the same sample areas as the P sample. After matching with the census lists and reconciliation, the P sample provides information about the population missed in the census, whereas the E sample provides information about erroneous census inclusions. This information is used in different ways to estimate the net coverage.

For 2010, we used logistic regression modeling to estimate the parameters in the DSE formula for census Data-defined (DD) status (whether an individual's data were collected as opposed to

imputed), Correct Enumeration (CE) and Match probabilities. We then estimate net coverage by comparing the estimate of the true population (from the DSE) to the census count, resulting in either a net undercount or a net overcount. The DSE for persons can be expressed as

$$DSE = \sum_{j \in C} \pi_{dd(j)} \times \frac{\pi_{ce(j)}}{\pi_{m(j)}} \times CB_j$$

With respect to the given estimation domain C, the modeled CE, Match and DD probabilities for census case j ($\pi_{ce(j)}$, $\pi_{m(j)}$, $\pi_{dd(j)}$) are obtained through logistic regression modeling. CB_j is an adjustment for correlation bias applied only for person estimation and described in the next section. DSEs for housing unit estimation contain only the CE and Match rate terms.

2.2 *Correlation Bias Adjustment*

In addition to the DD, CE, and Match rates, population estimates reflect a correlation bias adjustment that is applied to adult males only. It is estimated from sex ratios derived from demographic analysis, using separate counts for Blacks-alone-or-in-combination and Non-Blacks. As a result of this adjustment, the final DSE sex ratio will equal the demographic analysis sex ratio within each adult age-race group. This adjustment is applied after logistic regression modeling and does not impact the model diagnostics presented in this report. For more information and results of the correlation bias adjustment see Konicki (2012).

2.3 *Statistical Testing*

Comparisons made or implied for estimates in this report are statistically significant at the 90% confidence level ($\alpha = 0.10$) using a two-sided test. “Statistically significant” means that the difference is not likely due to random chance alone. Some comparisons among a limited set of modeling possibilities require that a choice be made, without implying statistical significance compared to competing alternatives.

2.4 *Software*

SAS Proc Logistic and Proc SurveyLogistic

We needed to use a standard software package that was available and easily usable by all programming partners. The Logistic and SurveyLogistic procedures fit the same model to a given dataset, but offer different features, each of which plays a role in the modeling process. Proc SurveyLogistic automatically calculates parameter standard errors that incorporate the complex sample design.

Proc Logistic calculates parameter estimates more quickly than Proc SurveyLogistic and offers additional features useful in the programming process.

Proc Gplot was also used to graph residual visual plots described below.

2.5 *Model Fit (risk function) by Log-Likelihood*

For any given choice of covariates, log-likelihood (which is always negative) is maximized to estimate parameters. In this document, log-likelihoods have their signs reversed to express as absolute values. Generally, a lower absolute log-likelihood represents a better fit, although when models with different numbers of parameters are compared, the possibility that the larger model is over-parameterized (i.e., the additional parameters are just fitting random variation) needs to be checked. All log-likelihoods in this document reflect sampling weights.

2.6 *Parameter Choice by Wald Likelihood tests*

Statistical significance of parameter estimates, as calculated by SAS Proc SurveyLogistic, will be a primary tool in selecting model variables. Proc SurveyLogistic generates an estimate of each parameter with its standard error estimate to determine statistical significance. Categorical variables with multiple degrees of freedom can be tested simultaneously from a separate table of Wald tests, which SAS displays along with a measure of its chi-square probability.

2.7 *Testing for Over-Parameterization by Cross-validation*

Cross-validation is a replication technique that adjusts the model fit measure for the possibility of over-parameterization, by creating test sets of the sample under which fit measures do not use the same observations used to construct the test parameterization. Because it is cumbersome to run in SAS (requiring a logistic regression run for every replicate), it is used sparingly to verify the final model selections against the most reasonable alternatives, not for every conceivable model. It is described in Mule and Olson (2005). Like the log-likelihood, its results are negative numbers which are expressed in this document as absolute values, with results closer to zero representing better fit.

2.8 *Continuous Variable Transformations by Residual Examination and Plots*

Logistic regression uses a logit link¹ to map the real line into the (0,1) interval for estimating probabilities. Since continuous variables cannot be assumed to invoke a logit effect on the outcomes on which they operate, some kind of transformation is often required to fit the curve. For purposes of CCM, the transformation is evaluated both visually and numerically using a fit measure, from among a limited set of reasonable alternatives.

3. **Limitations**

The data in this report have certain limitations to be noted when reading this document.

3.1 *Sampling Error*

Since the CCM estimates are based on a sample survey, they are subject to sampling error. As a result, the sample estimates will differ from what would have been obtained if all housing unit

¹ Although logistic regression can use other links, logit is most common and used in CCM. Alternatives were researched in Olson (2010).

persons had been included in the survey. The standard errors provided with the data reflect mainly variations due to sampling and they do not in general account for nonsampling errors which can be the principal source of error for very small geographic areas. Thus, the standard errors provide an indication of the minimum amount of possible error present in the estimates. See the forthcoming methodology report for more details on the variance estimation.

3.2 Nonsampling Error

Nonsampling error is a catch-all term for errors that are not a function of selecting a sample. They include errors that may occur during survey data collection and processing. For example, while an interview is in progress, the respondent may make an error in answering a question, or the interviewer may make an error in asking a question or recording the answer. Sometimes interviews do not take place or households provide incomplete data. Other examples of nonsampling error for the 2010 CCM include modeling error, synthetic error, and classification error. Unlike sampling error, nonsampling error is difficult to quantify.

3.3 Prefer to Use Same Covariates in Each Rate Model

It is certainly possible that a covariate useful in fitting one or two of the rate elements of the DSE is not necessary to include in the model of all three. However, research from the 2000 Accuracy and Coverage Evaluation (A.C.E.) had determined that the use of different characteristics for the CE and Match rates could create extreme variability in synthetic estimates (U.S. Census Bureau 2003). Therefore, to reduce this risk, the same characteristics were used in CE and Match rate modeling. We allowed for interactive effects to differ, if determined by fit measures to be significant. Because the DD model is based on such a large input data set, many interactions' tests were statistically significant that were not for the E and P samples. Interactions for DD were selected by the practical effect their size would impose on the DSEs, and the need to balance the effects measurable in the other models.

3.4 Standard Techniques and Software

The production environment limits the choices of modeling tools and techniques to those that are widely available and understood in the statistical community. The primary statistical modeling software is SAS, from which all diagnostic measures were produced. Some alternatives were investigated, but not deemed sufficiently better to use.

3.5 Modeling Assumption about Additivity

The primary reason to use statistical modeling in place of post-stratification is to reduce the number of required interactions, hence allowing the inclusion of more main effects. Regression modeling substitutes an assumption based on a linking function for interactions that are not specified. The models used in CCM use the logit link, which is most common in logistic regression.

3.6 Characteristics

For net coverage modeling, only characteristics available for each census and P-sample individual (person or housing unit) were used (for the reasons stated in Section 3.3), which limits

the usable covariates to those that are observed directly in both the census and Independent Listing or Person Interview (operations performed on P-sample members), or can be assigned through geography (such as tract rate measures).

4. Discussion of Results for Person Estimation

The CCM person model used the following characteristics from the 2000 A.C.E., with some changes to them (see Attachment 1 for more details). For presentation of software outputs in later sections, each variable has a six-letter abbreviation.

Table 1: Modeling Variables from the Census 2000 A.C.E.

Description	Abbrev.	2000 A.C.E.	2010 CCM
Race/Hispanic Origin Domain	Domain	7 groups	Same
Tenure: Owner or Renter	Tenure	2 groups	Same
Sex and Age	AgeSex	7-8 groups	9 groups
Tract Return Rate	Par_Rt	2 groups	Continuous*
MSA size and TEA ¹	MsaTea	4 groups	7 groups
Region	Region	4 groups	Same

* CCM uses Tract Participation Rate instead of Return Rate (see explanation in Attachment 1)

¹ Metropolitan Statistical Area size crossed with Type of Enumeration Area

Additional variables, whose use is made possible because of regression modeling, are

- Relationship to Householder (*HH_Rel*, 3 groups): Nuclear family member, adult child, other relative or non-relative
- Presence of Spouse in Household (*Spouse*, 2 groups): Whether the household contained a member with relationship “spouse”
- Replacement Mailing Area (*RpMail*, 3 groups): Blanketed, Targeted, Not
- Bilingual Area (*Biling*, 2 groups): Bilingual or Not

4.1 Are All These Characteristics of Value?

To check that all of the characteristics contribute to modeling, we ran a model using two-way interactions for the Race/Origin Domain, Age/Sex, and Tenure variables, plus main effects of other variables through SAS Proc SurveyLogistic, shown in Tables 2 and 3. [For this first analysis, Tract Participation Rate is modeled as a continuous variable with no transformation. See below for diagnostics related to the decision to model this as a continuous variable with no transformation.]

Table 2: Significance Test of Candidate Person Covariates: E Sample

Characteristic	DF	Wald Chi- Square	Pr > ChiSq
Domain	6	13.5692	0.0348
Tenure	1	21.8775	<.0001
AgeSex	8	22.8535	0.0036
Domain*Tenure	6	27.3380	0.0001
Domain*AgeSex	48	123.3918	<.0001
Tenure*AgeSex	8	53.9228	<.0001
Par_Rt	1	75.3434	<.0001
Region	3	9.6432	0.0219
MsaTea	6	12.6141	0.0496
Spouse	1	123.8769	<.0001
HH_Rel	2	573.7486	<.0001
RpMail	2	33.2881	<.0001
Biling	1	2.4238	0.1195

Every characteristic except Bilingual Area (Biling) is significant ($p < 0.05$).

Table 3: Significance Test of Candidate Person Covariates: P Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Domain	6	33.5554	<.0001
Tenure	1	52.5767	<.0001
AgeSex	8	97.0940	<.0001
Domain*Tenure	6	11.2292	0.0815
Domain*AgeSex	48	88.0743	0.0004
Tenure*AgeseX	8	65.2024	<.0001
Prt_Rt	1	27.4413	<.0001
Region	3	41.1752	<.0001
MsaTea	6	37.0831	<.0001
Spouse	1	209.9673	<.0001
HH_rel	2	624.2358	<.0001
RpMail	2	23.6329	<.0001
Biling	1	2.6405	0.1042

Again, Bilingual Area is non-significant.

Bilingual Area was shown to be significant after crossing it with the Participation Rate, and running the same model but with the new interaction:

P Sample:

	Param	SE	Wald Chi-Square	Pr > Chisq
Biling	0.3198	0.1716	3.4746	0.0623
Pr_Rt*biling	-0.5219	0.2456	4.5166	0.0336

Treated as a two-way interaction, the pair of effects were significant for the P sample. We accepted it for the E sample for consistency.

4.2 Changes to the Existing Variables

Should Age/Sex be expanded to nine categories from seven?

To test this, we parameterized Age/Sex using the seven categories, and then created an additional variable that parameterized the two additional categories. We ran this through the Proc SurveyLogistic to jointly test the significance of the two additional parameters.

	Effect	DF	Chi-Square	Pr > ChiSq
E Sample	Seven Levels	6	218.5137	<.0001
	Additional Two	2	12.6569	0.0018
P Sample	Seven Levels	6	485.7600	<.0001
	Additional Two	2	44.2409	<.0001

The additional two degrees of freedom are significant, and the nine categories will be used.

Should MSA/TEA be expanded to seven categories from four?

Similar to the above, we parameterized the four-level crossing of the Metropolitan Statistical Area size with Type of Enumeration Area (MSA/TEA), and then created an additional variable parameterizing the three additional categories:

	Effect	DF	Chi-Square	Pr > ChiSq
E Sample	MT-four_levels	3	11.8761	0.0078
	MT-seven	3	5.1200	0.1632
P Sample	MT-four_levels	3	15.0820	0.0017
	MT-seven	3	16.1282	0.0011

Although the new categories were not significant for the E sample, the three new levels are significant for the P sample. We accepted for use in both samples for consistency.

Should Participation Rate be modeled as continuous?

To test whether Participation Rate should be modeled continuously, we parameterized a Hi/Lo indicator for comparing each tract's rate to the national average of 74%. We then created a variable for the residual, equal to the difference between each observation's rate and the mean value of all rates with the same indicator value (mean low rate=65%, mean high rate=80%). Then we put both the indicator and continuous residual into the model:

	Effect	DF	Chi-Square	Pr > ChiSq
E Sample	Hi/Lo Indicator	1	38.9648	<.0001
	Continuous Residual	1	65.3615	<.0001
P Sample	Hi/Lo Indicator	1	13.1204	0.0003
	Continuous Residual	1	19.6656	<.0001

In both cases, the continuous version contributes significantly compared to just using the Hi/Lo indicator.

4.3 Transformation of the Participation Rate

Since the Participation Rate is a continuous variable, we explored whether it should be transformed. To test this, we ran the rate linearly (no transformation), squared, and as a square root, to see if it fits best using a transformation, using an exploratory model with the full interaction of Race/Origin Domain, Age/Sex and Tenure, plus main effects of all the other modeling variables. These three alternatives do not represent a comprehensive set of possible transformations, but can indicate whether the slope of the modeling covariate changes in different parts of its range.

Table 4: Fit Measures (in absolute value) for Transformations of Participation Rate

Participation Rate Transformation	Intercept Only	With Covariates
E Sample		
Square Root	84,616,602	81,133,116
Linear	84,616,602	81,134,356
Squared	84,616,602	81,136,559
P Sample		
Square Root	86,083,514	80,699,776
Linear	86,083,514	80,697,636
Squared	86,083,514	80,697,812

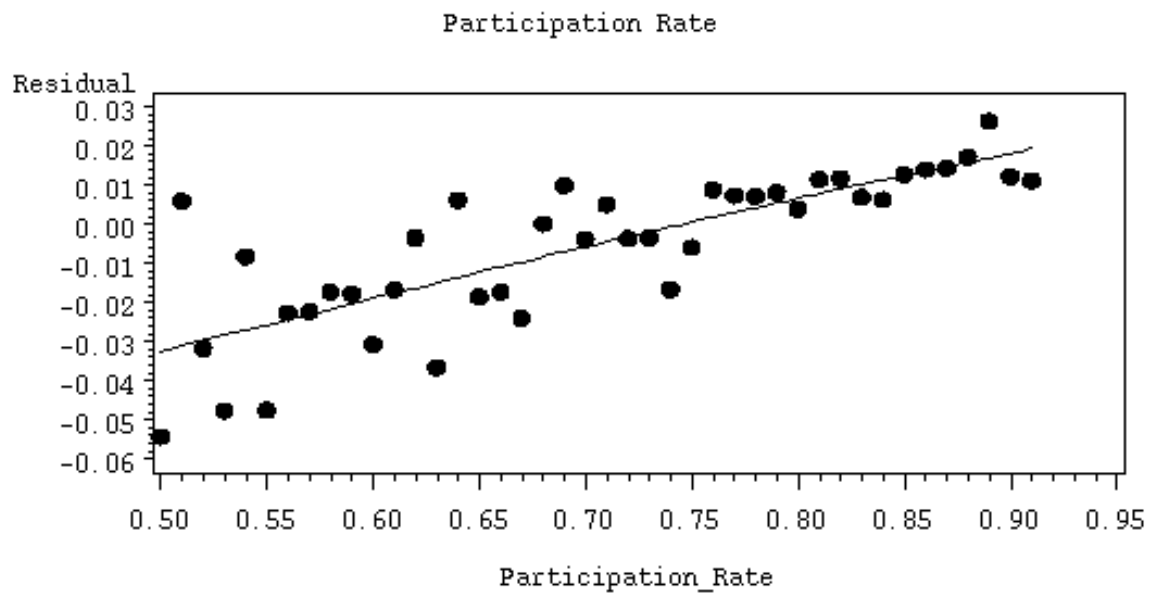
In Table 4, for the E sample squared appears worst and, for the P sample square root appears worst. Linear is indicated as the best choice, although the comparison does not imply statistical significance. This can also be examined visually using a residual plot.

We modeled each sample observation as if its Participation Rate was the national average. The residual was defined as the difference between a) the observed weighted CE (or Match) rate of all persons with the same integer percent Participation Rate; and b) the average predicted CE (or Match) rate using the logistic regression model, but assuming the Participation Rate for each person is equal to the national average Participation Rate.

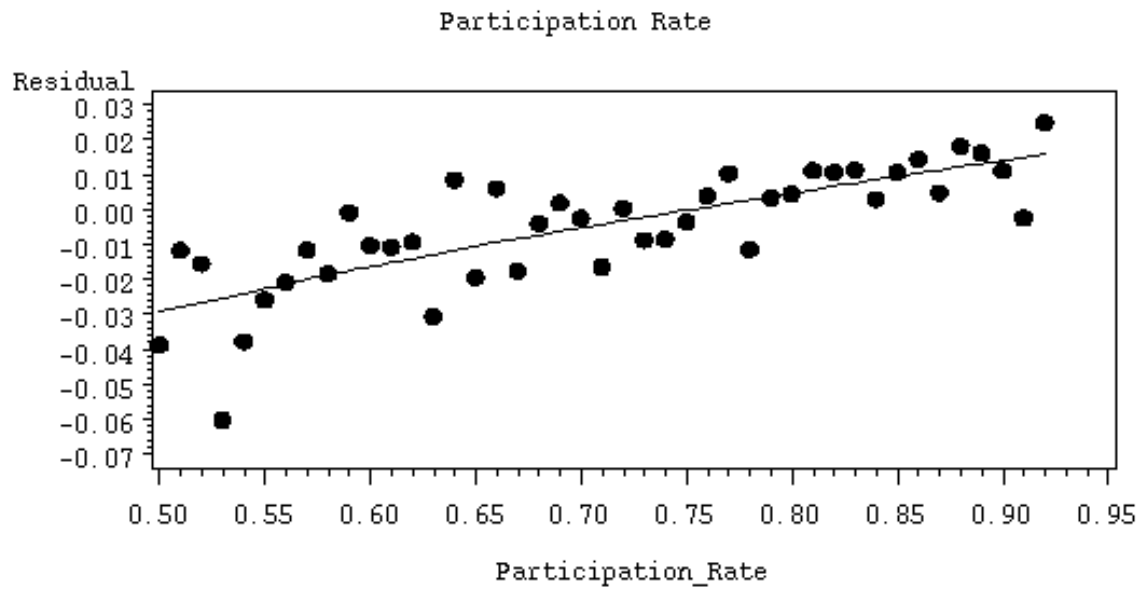
Looking at Plots 1 and 2 on the next page, both the CE and Match rate residuals appear to be a linear function of the Participation Rate, indicating no transformation was necessary.

We accepted the variables described in this section for inclusion in the CCM models.

Plot 1: CE Rate Residuals



Plot 2: Match Rate Residuals



4.4 Choice of Interactions

The following interactions were chosen based primarily on the strength of their statistical tests, but also with a view to consistency between the samples. The model was run through Proc SurveyLogistic to get Wald tests for the individual category sets, shown in Tables 5 and 6.

Table 5: Wald test for Significance of CCM Production Person Model Covariates: E Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Domain	6	7.6283	0.2666
Tenure	1	5.3909	0.0202
AgeSex	8	23.0232	0.0033
Domain*Tenure	6	11.4786	0.0747
Domain*AgeSex	48	129.8711	<.0001
Tenure*AgeSex	8	17.8463	0.0224
Domain*Tenure*AgeSex	48	75.2958	0.0071
Region	3	0.9836	0.8052
MsaTea	6	5.5423	0.4764
Biling	1	0.5332	0.4653
RpMail	2	29.9094	<.0001
Prt_Rt	1	18.8818	<.0001
HH_Rel	2	8.9467	0.0114
Spouse	1	37.7128	<.0001
Domain*Region	18	30.4659	0.0332
Prt_Rt*Tenure	1	14.3750	0.0001
Prt_Rt*Region	3	8.2329	0.0414
Region*Msatea	18	29.8671	0.0388
Tenure*Spouse	1	35.1305	<.0001
AgeSex*Spouse	8	134.3080	<.0001
rel_type*Spouse	2	53.3505	<.0001
Tenure*HH_Rel	2	10.0786	0.0065
AgeSex*Region	24	47.4355	0.0030
Prt_Rt*Biling	1	0.8121	0.3675
Prt_Rt*HH_Rel	2	37.3581	<.0001

All the interactions in the set are significant, except for interaction of Participation Rate with Bilingual, which was included for consistency with the P sample. The main effect for Region appeared non-significant, but must be considered together with the Age/Sex-Region interaction, which is highly significant. Hence, Region also was included as a main effect.

Table 6: Wald test for Significance of CCM Production Person Model Covariates: P Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Domain	6	28.6812	<.0001
Tenure	1	0.5879	0.4432
AgeSex	8	77.5740	<.0001
Domain*Tenure	6	12.1477	0.0588
Domain*AgeSex	48	87.8737	0.0004
Tenure*AgeSex	8	11.6513	0.1675
Domain*Tenure*AgeSex	48	79.8117	0.0027
Region	3	4.0209	0.2592
Msatea	6	41.3382	<.0001
Biling	1	5.0003	0.0253
RpMail	2	20.7896	<.0001
Prt_Rt	1	27.6655	<.0001
HH_Rel	2	4.3240	0.1151
Spouse	1	49.6793	<.0001
Domain*Region	18	29.0805	0.0474
Prt_Rt*Tenure	1	1.0032	0.3165
Prt_Rt*Region	3	4.9306	0.1769
Region*MsaTea	18	23.7397	0.1637
Tenure*Spouse	1	8.9085	0.0028
AgeSex*Spouse	8	119.7634	<.0001
HH_Rel*Spouse	2	94.2636	<.0001
Tenure*HH_Rel	2	32.3612	<.0001
Region*AgeSex	24	37.1140	0.0426
Prt_Rt*Biling	1	6.1807	0.0129
Prt_Rt*HH_Rel	2	4.5414	0.1032

For the P sample, three Participation Rate interactions were not individually significant, those with Tenure, Region, and Household Relationship. The Bilingual interaction is significant. Since the E and P models seem to prefer different Participation Rate interactions, we accepted all of them and kept this model. The model in Tables 5 and 6 was selected as the CCM Production Model for persons.

4.5 Choice of Model for the Data-Defined Rate

Since the same set of main effects and interactions were selected for the CE rate model and the Match rate model, we decided to use these same main effects and interactions for the DD model. Because of the large number of observations (300 million), every term was significant.

4.6 Testing for Model Fit

We tested the overall fit of the whole model by running a 20-replicate cross validation comparing models with smaller and larger numbers of covariates. Cross-validation tests for over-parameterization by testing the fit of each sample observation under a model that doesn't use the observation being tested, implicitly creating a penalty for increased parameter count. We present in Table 7 three models for illustration of the technique, although many more were tried during the model development process:

- the initial analysis parameter-checking model with two-way interactions of the Race/Origin Domain, Age/Sex, and Tenure variables and main effects only for all the other variables, used in Section 4.1 (Smaller model)
- the CCM production model
- the CCM production model with the two-way interaction of Tenure-MSA/TEA added.

Table 7: Cross-Validation Comparison of Candidate Models (in absolute value)

Model	Log Likelihood	Cross-Validation
E Sample		
Smaller	81,401,850	81,650,060
CCM	81,134,255	81,523,196
Larger	81,127,467	81,536,999
P Sample		
Smaller	80,993,642	81,267,932
CCM	80,698,913	81,222,180
Larger	80,694,788	81,254,553

As it mathematically must, each larger model reduces (improves) the absolute log-likelihood from the smaller one. But the larger absolute cross-validation for the “Larger” model confirms it would be an overfit.

4.7 Final Model Check

It is possible for a model that fits overall to fail in some parts of it. To examine this, we divided the fitted model into groups defined by five-percent ranges of modeled CE (and Match) rates, with a minimum 250 observations in each group, in Tables 8 and 9. Since modeled CE (and Match) rates are calculated to many decimal places, no modeled rate ever exactly equals the boundary value of a range. Standard errors reflect only the sampling variance of the observations, not the variance among estimated rates within the categories.

Table 8: CE Rate Model Fit Check for E-sample Persons

Modeled Range	Sample Count	Mean Value		Standard Error	
		Modeled	Observed	Modeled	Observed
< 75.0	391	72.62	76.73	2.51	3.92
75.0 - 80.0	4,807	78.64	80.20	0.64	0.91
80.0 - 85.0	32,765	83.03	83.57	0.30	0.34
85.0 - 90.0	93,984	87.89	87.81	0.17	0.20
90.0 - 95.0	148,144	92.57	92.30	0.10	0.13
95.0 - 100.	103,446	96.06	96.26	0.09	0.11

Table 9: Match Rate Model Fit Check for P-sample Persons

Modeled Range	Sample Count	Mean Value		Standard Error	
		Modeled	Observed	Modeled	Observed
<65.0	667	61.63	63.24	2.77	4.93
65.0 - 70.0	2,316	68.03	67.30	1.07	1.60
70.0 - 75.0	7,109	72.97	74.03	0.64	0.76
75.0 - 80.0	18,335	77.88	77.67	0.38	0.47
80.0 - 85.0	39,567	82.81	82.96	0.25	0.32
85.0 - 90.0	73,166	87.76	87.75	0.17	0.21
90.0 - 95.0	117,058	92.84	92.73	0.12	0.14
95.0 - 100.0	97,594	96.19	96.25	0.09	0.11

None of the row discrepancies are statistically significant. The lowest modeled CE rates are lower than one would like to see, but they represent only about 0.15% of the sample cases. The lowest modeled Match rates are just slightly lower than their observed rates, implying that no major distortions would be caused by applying the modeled rates as reciprocals in DSEs.

This verification concludes that this model is acceptable for CCM production use.

5. Discussion of Results for Housing Unit Estimation

The census collects very few characteristics about housing units, so housing unit modeling relies primarily on characteristics of the occupants (if any), the neighborhood in which the unit is located, and census operational measurements. Additionally, the CE and Match rates for housing units are very high, 97.3% and 97.0% respectively, with their negative outcomes highly clustered among the primarily sampling units. Housing unit modeling involved many fewer covariates than person modeling.

The characteristics used in housing unit modeling are (along with abbreviated names used in output tables, for those not also used in person modeling)

- Occupancy and occupants: Four categories intersecting Owner or Renter with an indicator of whether the householder was non-Hispanic White, plus a fifth category for Vacant units [*OcTnNH*]
- Structure Size and Type: Four categories for Single Units, Small (2 to 9) Multi-Units, Large (10+) Multi-Units, and Trailers [*Struct*]
- Geographic characteristics: Region (4 categories) and MSA/TEA (7 categories) are defined the same way as in person modeling.
- Replacement Mailing and Bilingual Area: Areas of enhanced census enumeration efforts in areas expected to be difficult to count. Defined as in person modeling.
- Census address list building rates: The percent of final census enumerated housing units in the collection tract that had been included on census address lists as of two phases of address list building, Address Canvassing and Enumeration (which corresponds to the list of housing units to which census forms were distributed). A low rate of these measures indicates that the neighborhood was experiencing new construction or was difficult to count. [*AdCn_rt* and *Enum_rt*]

These characteristics are described more fully in Attachment 1: Variable Definitions.

5.1 *Are All These Characteristics of Value?*

A main effects model is shown in Tables 10 and 11 to check the contribution of each characteristic.

Table 10: Significance Test of Candidate Housing Unit Covariates: E Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Struct	3	585.5828	<.0001
Region	3	17.8393	0.0005
MsaTea	6	96.3511	<.0001
OcTnNH	4	540.8360	<.0001
Enum_rt	1	10.3157	0.0013
Adcn_rt	1	4.5937	0.0321
RpMail	2	23.8872	<.0001
Biling	1	0.5871	0.4435

Table 11: Significance Test of Candidate Housing Unit Covariates: P Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Struct	3	104.7728	<.0001
Region	3	0.2857	0.9627
MsaTea	6	19.2664	0.0037
OcTnNH	4	654.9083	<.0001
Enum_rt	1	3.2076	0.0733
Adcn_rt	1	20.5557	<.0001
RpMail	2	3.2298	0.1989
Biling	1	2.9707	0.0848

Every characteristic was significant ($p < 0.05$) in the E Sample, except Bilingual Area. In the P Sample, Region, Enumeration List Rate and Replacement Mailing status did not test as significant. However, all were carried forward into the modeling efforts.

5.2 *Transformation of Continuous Variables*

A continuous variable sometimes requires a transformation. Plots of residuals for the CE and Match rates of the two address list building rates are presented in Attachment 2. All appear visually to increase in slope as the modeled rate increases, implying that the square root should not be expected to fit well. To check for reasonable transformations, the two rates were also modeled as a square root, linear, and squared transformation, in Table 12. Either a squared or linear transformation might fit better in any particular case, depending on whether the increase in slope is sufficient.

Table 12: Log Likelihood Rates (in absolute value) of Model Fits Testing for Best Transformation

Transformation	Address Canvassing Rate	Enumeration List Rate
CE Rate		
Square root	29,597,158	29,687,037
Linear	29,594,534	29,683,421
Squared	29,589,337	29,676,127
Match Rate		
Square root	31,270,724	31,271,031
Linear	31,269,440	31,268,389
Squared	31,266,917	31,263,272

In each case, the squared transformation fits best (has lowest absolute log-likelihood measure), and square root fits worst, although statistical significance is not implied. Therefore, squared transformations were applied to all models.

5.3 Choice of Interactions

Housing unit modeling cannot support as extensive a set of interactions as was used in person modeling. Wald tests supported only three interactions in each model, representing 19 additional degrees of freedom, shown in Tables 13 and 14. The two samples preferred different address list building rates to be interacted with the Structure Size/Type characteristic. This decision is discussed in the next section.

Table 13: Wald test for Significance of CCM Production Housing Unit Model Covariates: E Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Struct	3	8.1919	0.0422
Region	3	12.8801	0.0049
MsaTea	6	39.5536	<.0001
OcTnNH	4	5.9041	0.2064
Enum_rt^2	1	12.4405	0.0004
AdCn_rt^2	1	24.3893	<.0001
RpMail	2	11.2364	0.0036
Biling	1	0.2049	0.6508
OcTnNH*AdCn_rt^2	4	34.1847	<.0001
Struct*OcTnNH	12	124.9201	<.0001
Struct*Enum_rt^2	3	16.4401	0.0009

Table 14: Wald test for Significance of CCM Production Housing Unit Model Covariates: P Sample

Characteristic	DF	Wald Chi-Square	Pr > ChiSq
Struct	3	6.0621	0.1086
Region	3	0.2417	0.9706
MsaTea	6	18.7113	0.0047
OcTnNH	4	1.8870	0.7565
Enum_rt^2	1	10.1145	0.0015
AdCn_rt^2	1	18.1808	<.0001
RpMail	2	1.7885	0.4089
Biling	1	3.2308	0.0723
OcTnNH*AdCn_rt^2	4	13.4267	0.0094
Struct*OcTnNH	12	22.0940	0.0365
Struct*AdCn_rt^2	3	11.2324	0.0105

5.4 Testing for Model Fit

A cross-validation with 20 replicates tested four candidate models under each sample. The candidate models involved all the main effects, plus the two interactions that had been deemed to fit both samples. The four candidates represent choices of interaction between the two address list building rates with the Structure Size/Type covariate, shown in Table 15:

Table 15: Model Fit Measures (in absolute value) among Rate Interactions with Structure Type

Interact Structure Size/Type with...	Correct Enumeration		Match	
	LogLikelihood	Cross-Validation	LogLikelihood	Cross-Validation
Neither	29,077,644	29,551,304	31,044,517	31,552,125
Address Canvassing	29,039,511	29,553,926	30,984,172	31,539,599
Enumeration List	29,004,912	29,508,772	30,998,042	31,592,526
Both Rates	28,970,004	29,513,325	30,969,269	31,590,196

The CE model diagnostics preferred to include only the Enumeration List rate interaction (it had the smallest absolute cross-validation measure); adding the Address Canvassing rate interaction did not improve it. Match rate modeling distinctly preferred using only the Address Canvassing rate interaction, as cross-validation measures got much worse when Enumeration List terms were added. So, different interactions for these terms were used in the two models.

5.5 Final Model Check

As we did for persons, we ran a simple residual check to help determine if some parts of the range of modeled values were fit poorly, shown in Tables 16 and 17. Standard errors reflect only sampling variance, not variation among the modeled values within the categories.

Table 16: CE Rate Model Fit Check for E-sample Housing Units

Modeled Range	Sample Count	Average Rate		Standard Error	
		Modeled	Observed	Modeled	Observed
<75.0	425	72.29	79.48	2.77	3.66
75.0 – 77.5	430	76.40	78.38	1.87	3.87
77.5 – 80.0	714	78.95	80.23	1.27	2.69
80.0 – 82.5	1,394	81.42	84.30	0.90	1.55
82.5 – 85.0	1,541	83.86	82.26	0.73	2.00
85.0 – 87.5	2,575	86.21	83.40	0.67	2.67
87.5 – 90.0	2,960	88.91	90.40	0.55	1.06
90.0 – 92.5	6,167	91.39	91.61	0.58	0.63
92.5 – 95.0	15,422	93.94	92.92	0.40	0.73
95.0 – 97.5	35,066	96.47	96.77	0.24	0.25
97.5 – 100.0	106,223	99.06	99.06	0.06	0.07

Table 17: Match Rate Model Fit Check for P-sample Housing Units

Modeled Range	Sample Count	Average Rate		Standard Error	
		Modeled	Observed	Modeled	Observed
< 70.0	474	67.89	60.06	3.70	12.71
70.0 – 72.5	350	71.57	73.58	2.43	5.67
72.5 – 75.0	527	73.82	76.28	1.84	3.11
75.0 – 77.5	663	76.41	79.34	2.14	3.18
77.5 – 80.0	955	78.93	80.51	1.58	2.72
80.0 – 82.5	1,289	81.26	84.76	1.17	1.88
82.5 – 85.0	1,719	83.98	82.71	1.18	3.12
85.0 – 87.5	2,655	86.38	85.36	0.80	2.20
87.5 – 90.0	4,782	88.81	88.29	0.83	1.41
90.0 – 92.5	6,293	91.35	91.43	0.51	0.70
92.5 – 95.0	11,811	93.84	94.12	0.32	0.43
95.0 – 97.5	32,252	96.59	96.46	0.21	0.27
97.5 – 100.0	103,117	98.75	98.77	0.07	0.08

None of the differences is significant. The variation between the observed and modeled rates in the lowest category in each list was more than one would ideally like to see, but they represented fewer than 500 cases or 0.30% of the sample. Since modeled Match rates were applied in the DSE formula as a reciprocal, it's good to see the low rates were not underestimated.

These covariates in Section 5.3 are accepted as the final housing unit model.

6. Conclusions

The person model diagnostics show that it was appropriate to define the independent variable for person estimation using the same category levels as used in the 2000 A.C.E. for Race/Hispanic Origin Domain, Tenure, and Region. For Age and Sex, changing from seven or eight levels to nine, as well as changing MSA/TEA from four levels to seven is also appropriate. The Participation Rate variable was improved by modeling it as continuous, instead of using two categories as was done in 2000 A.C.E. with the Return Rate. All these characteristics, as well as a three-category Relationship to Householder variable, a two-category indicator for presence of Spouse, a three-category Replacement Mailing Area variable, and a two-category Bilingual Area variable, were determined to be of value using a Wald Chi-Square test. This test was also used to determine which interactions to include in the models. The analysis indicates that it is acceptable to use the same main effect variables and interactions in the models for E-sample CE, P-sample Match, and census DD rates.

The housing unit model captured a good amount of dispersion in the modeled rates. Given the inherent limitations due to its smaller effective sample size, very high observed CE and Match rates, and the number of characteristics available, the model captured low rates reasonably well. The large number of operational variables included in the model should serve the CCM function of providing a framework to study possible improvements to future censuses. The analysis indicated that the E-sample and P-sample models should use different address building rates to be interacted with the Structure Size/Type characteristic.

References

- Bentley, M. (2008), "Specifications for Bilingual Form Distribution in the 2010 Census (Phase 1)," DSSD 2010 Decennial Census Memorandum Series #B-4.
- Datta, A.R., Yan, T., Evans, D., Pedlow, S., Spencer, B., Bautista, R. (2012), "Final Report: 2010 Census Integrated Communications Program Evaluation (CICPE)," 2010 Census Planning Memoranda Series No. 167.
- Gilary, A. (2011), "Recursive Partitioning for Racial Classification Cells," U.S. Census Bureau Research Report Series (Statistics #2011-04).
- Haines, D. (1999), "Accuracy and Coverage Evaluation Survey: Logistic Regression Modeling for Poststratification Variable Selection," DSSD Census 2000 Procedures and Operations Memorandum Series Q-6.
- Konicki, S. (2012), 2010 Census Coverage Measurement Estimation Report: Adjustment for Correlation Bias," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-11.
- Letourneau, E. (2010), "Specification to Identify Replacement Mailing Housing Units in the 2010 Census," DSSD 2010 Decennial Census Memorandum Series #G-04-R1.
- Mule, T. and Olson, D. (2005), "Initial Results of Preliminary Net Error Empirical Research Using Logistic Regression," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-03.
- Mule, T. (2008), "2010 Census Coverage Measurement Estimation Methodology," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18.
- Mulligan, J. and Davis, P. (2012), "2010 Census Coverage Measurement: Description of Race/Hispanic Origin Domain," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-45.
- Olson, D. and Springer, M. (2008), "2006 Census Coverage Measurement: Net Error Modeling Pseudo-Estimates," DSSD 2010 Census Coverage Measurement Memorandum Series #2006-E-09.
- Olson, D. (2010), "Research of Alternative Linking Functions," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-23.
- Rothhaas, C., Bentley, M., Hill, J. M., and Lestina, F. (2011), "2010 Census: Bilingual Questionnaire Assessment Report," DSSD 2010 CPEX Memorandum Series #C-01.
- U.S. Census Bureau (2003), "Technical Assessment of A.C.E. Revision II." March 12, 2003

U.S. Census Bureau (2004), "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology."

Ward, J. (2011), "Documentation for the 2010 Census Address Frame Combination File, version #1," DSSD 2010 Decennial Census Memorandum Series #D-19.

West, K., Robinson, J.G., and Bentley, M. (2005), "Did Proxy Respondents Cause Age Heaping in the Census 2000?" *Proceedings of the Section on Survey Research Methods*, pp. 3658-3665.

Attachment 1: Variable Definitions

Race/Origin Domain

The 2010 Census and CCM questionnaires provide for six major race categories, resulting in 63 possible combinations into which a respondent can identify; and when combined with two Hispanic origin responses, creates 126 possible categories. The Census 2000 A.C.E. modeled race and Hispanic origin using seven categories, called “Race/Origin Domains” into which each census and P-sample person was assigned to exactly one. This categorization is mainly for modeling; most publications present tables tabulating each census race classification either alone or in combination with others. It was considered during the CCM research cycle to expand the race modeling of multi-racial persons and of Hispanics (who in almost all cases are assigned to the Hispanic Domain, ignoring their race identification). However, research by Gilary (2011) could not find any specific improvements to recommend due to the complexity and number of multi-race combinations, high imputation rate for race among Hispanics, and inconsistency of race reporting between the census and P sample for both Hispanics and multi-race persons.

The 2010 CCM uses the same seven-category Race/Origin Domain variable from the Census 2000 A.C.E. It is described fully in Mulligan and Davis (2012).

Tenure

The distinction between the rates for Owners and Renters is one of the most important in coverage measurement modeling. Everyone in the household is modeled and tabulated as an Owner if any household member owns the housing unit (with or without a mortgage), including a boarder or roommate who pays rent to another household member.

Age and Sex

The original version of the 2000 A.C.E. partitioned the population into seven categories of age and sex, with all children 0-17 years old in one category and adult males and females categorized separately into age groups 18-29, 30-49 and 50-up. The A.C.E. Revision II split the children into two age categories from 0-9 and 10-17. The 2010 CCM further splits the younger category into 0-4 and 5-9 age groups. Statistical justification for this is tested in Section 4.2.

During much of the CCM developmental cycle, it had been planned to model age as a continuous variable using splines (Olson and Springer 2008). Questions about the technical application of this technique arose because of irregularities in the shape of the spline curve at adult ages divisible by 5, where it is believed that “heaping” due to round-off in proxy reports (those by a non-household member; West et. al. 2005) is correlated with reduced overall accuracy of those reports, and hence lower observed CE and Match rates. Because it is likely that the age of many such persons was not consistent in the census and the P sample, concerns arose about the naivety of the assumption that, for example, a 45-year-old census person should be assigned for estimation purposes the modeled rate of a 45-year-old P-sample member, if the two ages could not generally be assumed to have been reported consistently. Therefore, the CCM reverted to the use of traditional age and sex categories for modeling.

Region

The Census Bureau divides the country into four Regions – Northeast, Midwest, South and West. It had been briefly considered to expand geographic modeling to reflect the nine Census Divisions (which nest within Region), but was determined that so many categories could not support the number of expected interactions. Region (which models as three degrees of freedom) was ultimately interacted with four other variables totaling 21 degrees of freedom as main effects, creating a total of 63 interactions. Division could not have been interacted so broadly.

Participation Rate

The 2000 A.C.E. had post-stratified most population groups into those living in a Census Tract that reflected a high or low return rate (percent of occupied households who mailed back their census forms.) In 2010, the Census Bureau produced a continuously updated measure of Participation Rate, a very similar concept, that tracked the rate of form return continually during the census period. Because the final tallies of this value had already been compiled, it was used in modeling instead of return rate. The rate was bottom- and top-coded at 50% and 92%, due to sparsity of observations outside that range. It is discussed in Sections 4.2 and 4.3.

Metropolitan Statistical Area and Type of Enumeration Area (MSA/TEA)

The 2000 A.C.E. used a hybrid area characteristic for MSA size and the Mailout TEA. In 2010, MSAs were divided by size into Large (the largest 12 MSAs; over 4 million population), Medium (over 500,000), and Small. In 2000, MSA was crossed with Mailout into four categories, while 2010 uses seven categories: Mailout areas divided into Large, Medium, Small, and non-MSA areas; non-Mailout into MSA and non-MSA; and a separate category for the Update/Enumerate TEA. The covariates associated with the additional degrees of freedom are tested in Section 4.2.

New Modeling Variables

The above characteristics had been used in the 2000 A.C.E. post-stratification. Regression modeling allowed the possibility of expanding the number of characteristics that could be used.

Household Composition: Presence of Spouse and Relationship to Householder

It has long been attempted by the census coverage measurement program to define household composition characteristics that could be used in post-stratification or modeling. A.C.E. had experimented with (but ultimately did not use for post-stratification) a two-type classification that took household size and relationship structure into account (Haines 1999), and the A.C.E. Revision II used household size and nuclear family classifications (U.S. Census Bureau 2004).

For use in the 2010 CCM, household relationships and composition were defined by

- Presence of Spouse: a two-level indicator of whether the household contained a person with reported relationship of spouse (to the householder/reference person) or not

- Relationship to Householder: A three-level classification for nuclear family members, adult children, and other household members

Whether a household contains a spouse (i.e., is headed by a married couple) is one of the most significant predictors of whether it will mail its form back (Datta et al. 2012). Because households can change composition between Census Day and CCM Interview Day, the spousal status of some P-sample households could change. A P-sample person who is a nonmover or outmover was assigned the spousal status of the household as of Census Day, while an inmover was assigned the status from CCM Interview Day.

The Relationship to Householder categories are designed to reflect known differences in the census inclusion rates, while keeping the number of categories small. The final categories are

- Nuclear Family Member: Householder; Spouse; Child of householder (biological, step, or adopted) age 0-17
- Adult Child: Child of householder (biological, step, or adopted) age 18 or over
- All other household members.

These classifications correspond well to observed rate differences in the samples, with Adult Children showing rates between those of the other two categories, but closer to Other Household Members for CE and closer to Nuclear Family Members for Match.

Observed 2010 CCM rates for defined Householder Relationship categories

Relationship to Householder	Correct Enumeration	Match
Nuclear Family Members	93.01	92.40
Adult Children	87.77	89.78
Other Household Members	86.75	83.07

Operational Characteristics: Bilingual Questionnaire and Replacement Mailing Areas

The Bilingual Questionnaire and Replacement Mailing Distribution were large operations designed to enhance census response in areas expected to have low response without them. Bilingual (English and Spanish) census questionnaires were mailed to housing units in select areas that could require Spanish language assistance to complete their census form (Bentley 2008; Rothhaas et al. 2011). The Census Bureau also mailed a replacement mailing package to some housing units in Mailout/Mailback areas of the country that had low mail response in Census 2000. Areas with low response in Census 2000 had a blanketed distribution where all housing units received a replacement mailing. For areas with mid-range response in 2000, only nonresponding housing units received a replacement mailing; this is referred to as targeted distribution (Letourneau 2010). Since these operations were assigned to entire collection blocks, they can be used as modeling variables because all P-sample members can be associated with a collection block. (Operations that only visit individual housing units, such as nonresponse followup, cannot be assigned to the P sample and hence cannot be used in net coverage modeling.) All housing units and their residents in Targeted areas are modeled with that characteristic, regardless of whether the particular unit was sent a replacement form. These covariates are included to reduce synthetic bias when analyzing the results from the operations.

HOUSING UNIT MODELING VARIABLES

The census does not collect nearly as many characteristics about housing units as it does about persons, so fewer modeling characteristics are used.

Region, MSA/TEA, and Replacement and Bilingual Area status, are defined identically as in person modeling.

Occupancy and Occupant

Person modeling divides all census persons into Owners and Renters; housing units can have the additional characteristic of Vacant. For housing unit modeling, the former two were subdivided into occupancy by a non-Hispanic white, or any other householder. In 2000 A.C.E., race and Hispanic origin were modeled using six of the seven Race/Origin Domain categories from person estimation, but this was determined to be too many categories for use in 2010. A non-Hispanic white householder is one who reported a race of white alone, and did not indicate Hispanic origin. The Owner/Renter determination was based on the Census Day occupants when available, and Interview Day for P-sample occupants if necessary.

Structure Size and Type

Housing unit estimation has traditionally modeled three sizes of dwelling structures: Single Unit, Small Multi-Unit (2 to 9), and Large Multi-Unit (10 or more). In 2010, Trailers were added as a fourth category. The unit count of a structure for the P sample was directly observed and recorded by the enumerators. The count for the census and E sample had to be processed from address files, using an algorithm similar to that in Ward (2011).

Address Canvassing and Enumeration Rates

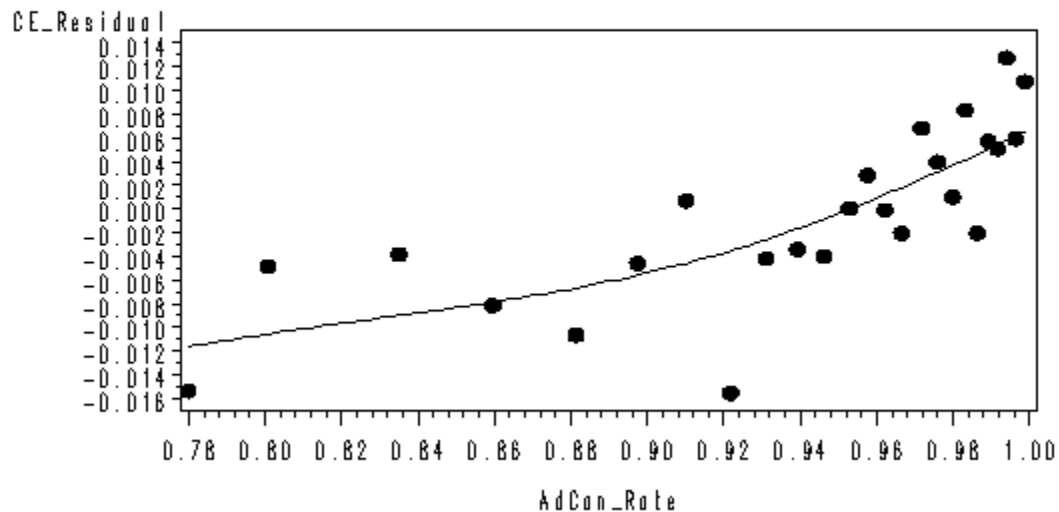
Because the census collects few characteristics about housing units, much of the modeling relies on neighborhood characteristics. The Address Canvassing and Enumeration lists are the names of two phases of address list construction that correspond to mileposts of census operations. Address Canvassing is the first phase of census address list building that reviews all the addresses to be counted that were known up to that point. The Enumeration list is the set of the known addresses to which census forms were distributed. Its name does *not* refer to the complete operation of counting all census persons. (Housing units added after the creation of the Enumeration list received forms in a separate operation, or were interviewed in person.) The rates used in modeling represent the fraction of housing units enumerated in the final census that were present on the appropriate list at the start of the corresponding operation. A low rate suggests that many units were added to the census roster around the time of the census, which usually reduces accuracy. The rates were tallied within collection tract. In tracts containing fewer than 100 census units, a weighted average was constructed between the actual observed rate of the tract and the national average rate, which was 91.65% for Address Canvassing and 97.76% for Enumeration. The Address Canvassing rate was bottom-coded at 78% and the Enumeration list rate at 89%, due to sparsity of data below those values.

Attachment 2: Housing Unit Rate Residual Plots

Curves fit using a cubic spline, with smoothness parameter chosen to show only major slope changes.

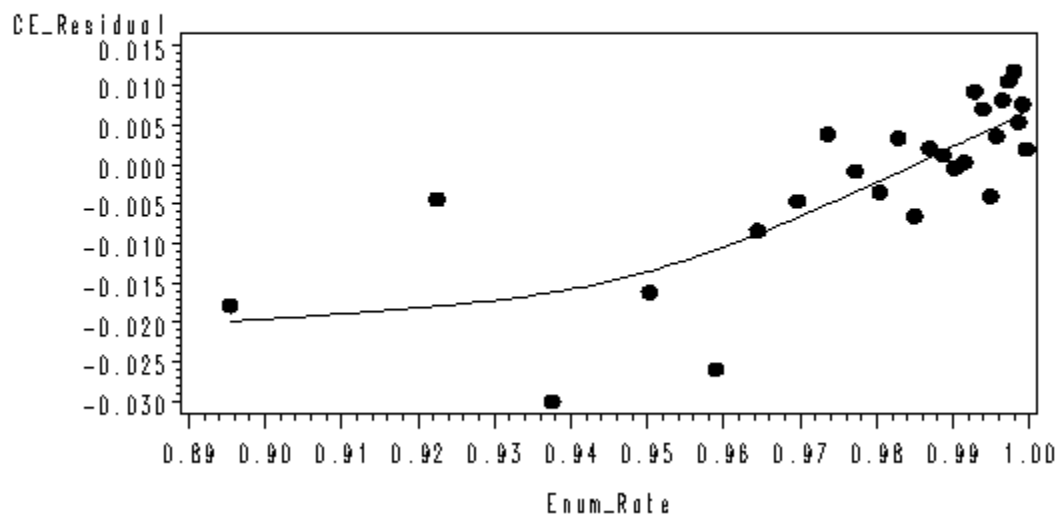
CE Rate Residuals for Housing Units

Address Canvassing Rate



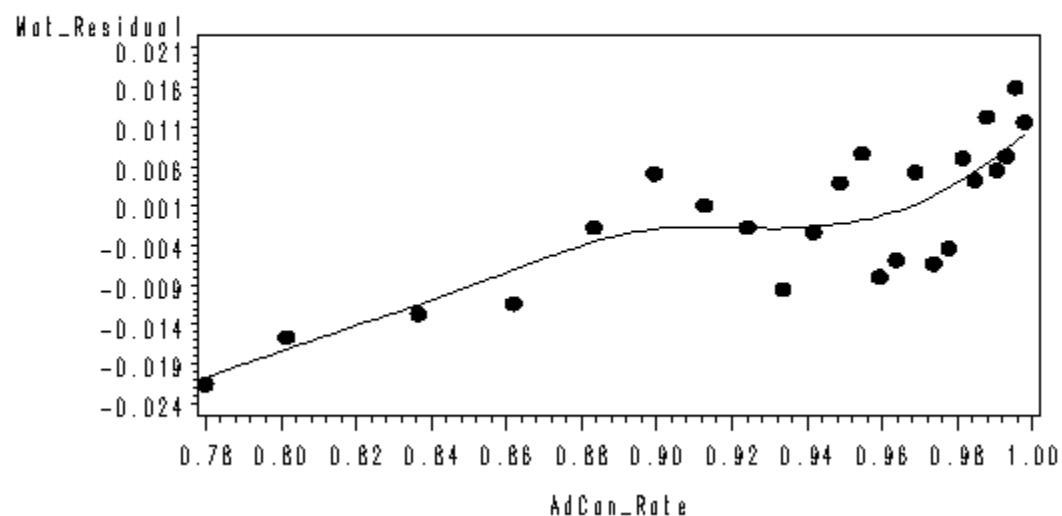
CE Rate Residuals for Housing Units

Enumeration List Rate



Match Rate Residuals for Housing Units

Address Canvassing Rate



Match Rate Residuals for Housing Units

Enumeration List Rate

